Weekly Report

2018.0611-2018.0618

1.This Week

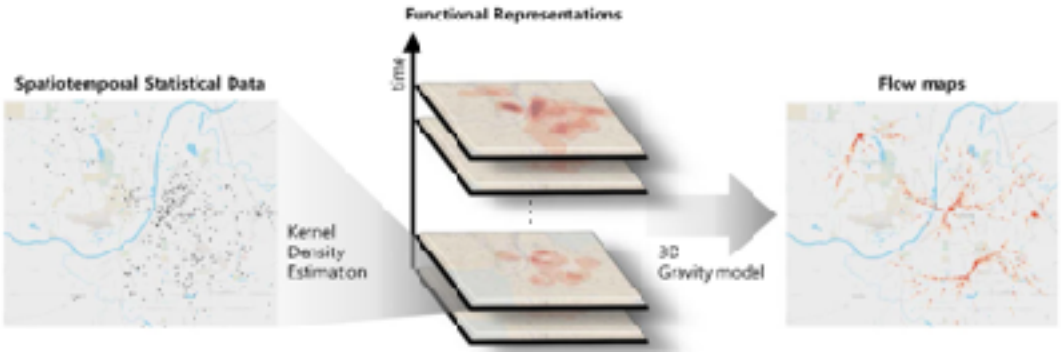
Power Grid Deep Learning Paper
1.Retrain the LSTM model with more accurate classification labels. (The label space is set to be the entire link and node set.)
2.Incremental Idea:
Since we need to deal with incomplete training data (data with only a few nodes' information in the grid), I am thinking of using the concept of active learning to select the most informative instances to do the training (training with less informative incomplete data may decrease the model accuracy). So I'm reviewing the query strategies of active learning last week (as shown in the reading report 1: active learning literature survey).
Things we still need to do: (a) define the meaning of "informative" in the case of power grid simulation data; (b) find a way to select the most informative instances out.

Power Grid Deep Learning Paper
1. Zongzhuang is realizing the CUSUM algorithm with the real case power grid data (data of 36 nodes in the power grid).

Others
1.review the Chinavis paper
2.learning basic theories (especially mathematical backgrounds) about machine learning and deep learning.

Paper Reading
1.Active Learning Literature Survey
This paper describes the basic idea and scenarios of active learning. My focus on reading this paper is the query strategies to find the most informative instance (to be labeled by the oracle). Common query strategies include uncertainty sampling, query-by-committee, expected model change, expected error reduction, variance reduction and density-weighted methods. Among these methods, uncertainty sampling is of the smallest complexity, but the result highly depends on the data characteristics. The following methods have, in general, an increasing accuracy but also increasing complexity.

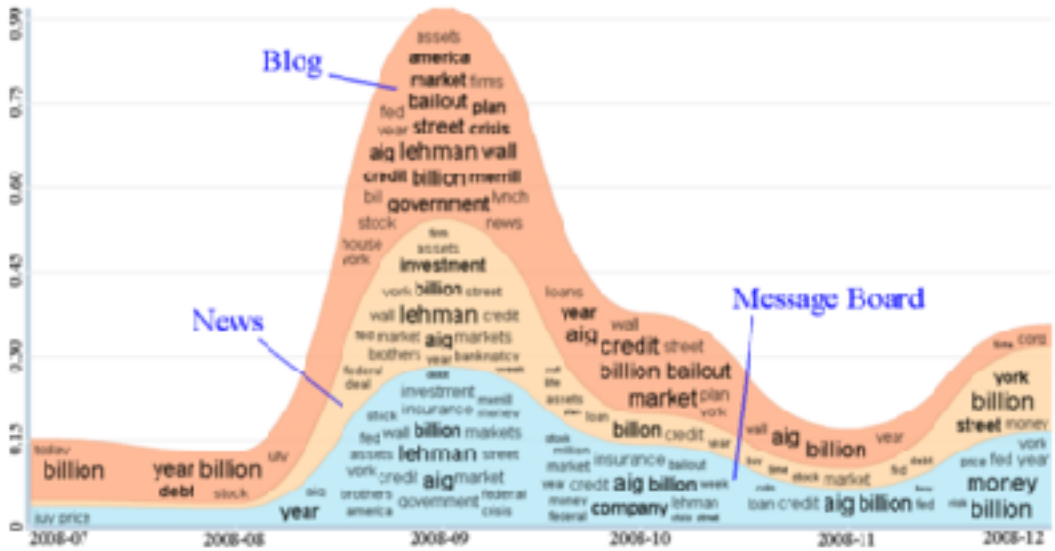
2. Deep Active Learning for Named Entity Recognition
This paper gives an example of how active learning can be applied to deep learning while assuring the computational efficiency. The classification model used



in this paper is a CNN-LSTM model, and the retraining process of active learning is carried out by mixing the newly annotated labels with the old ones and updating neural network weights for a small number of epochs. In this way, the retraining cost is reduced. The query strategy of this paper is uncertainty-based sampling. This is a typical paper of using active learning to reduce the label amounts, achieving 99% performance with only 25%-30% of labeled data.

3. Data Flow Analysis and Visualization for Spatiotemporal Statistical Data without Trajectory Information

This paper introduces an approach of extracting the flow fields from continuous data distributions rather then spatiotemporal trajectories. The main idea is to represent the discrete data as a continuous function (with KDE) and then extract the flow with a gravity model. This method can also be applied to our cell station data to analyze the population flow and may even be extended to simulate the trajectories of origin-destination data.



4. Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora

In the field of topic evolution mining, two major methods is evolutional clustering (as described in this paper) and dynamic topic model (as described in the next paper). This paper propose an approach to detect cluster evolution patterns from multiple correlated time-varying text corpora. It formulates the cluster mining process as a series of hierarchical Dirichlet processes by adding time dependencies to the adjacent epochs, and uses a cascaded Gibbs sampling scheme to infer the model. Evolving patterns like emerging, disappearing, evolution within a corpus and across different corpora can be found. (This paper is published on sig kdd, so the mathematical deduction takes the most part of the paper instead of visualization designs.)

5. Dynamic Topic Models

As mentioned in the former reading report, dynamic topic model is another typical approach for topic evolution mining. This paper develops sequential topic models for discrete data. The main idea is to apply Gaussian time series to logistic normal topic proportion models. The Gaussian time series on the natural parameters of the multinomial topics is used to represent the topics. And the approximate inference is carried out through Kalman filters and nonparametric wavelet regression.

2.Progress

Work	Deadline	Progress
Power grid paper with Deeping learning	-	Retraining the model
SQC Paper	-	About to started.